Experimental Design to Publication
Ulrich Dirnagl

Please visit the Brain & Brain PET 09’ website for downloading further information:
http://www2.kenes.com/brain/programme/Documents/Dirnagl_Course.pdf

Course outline

• Do we have a problem?
• The checklist
• Design: Internal validity
• Design: External validity
• Design: Statistics (focus on Power)
• Reporting
• SOPs
• GLP

Ischemic stroke in the rodent is a treatable disease!

A typical intervention in exp. stroke studies reduces infarct sizes by 30-50%.
Neuroregenerative strategies can improve functional outcome even after infarct maturation.

Only thrombolysis clinically effective!

I.v. thrombolysis is the only clinically proven pharmacological therapy in ischemic stroke. Thrombolysis can benefit less than 5% of all stroke victims.
There is no ‘neuroprotection’ in human stroke. There is no ‘neuroregeneration’ in human stroke.

The Problem

> 150 negative clinical Phase III studies, > 8 billion USD burnt. 2 positive trials: NINDS, ECASSIII

www.strokecenter.org
The Problem

A 'nuclear winter' for neuroprotection?

The Problem

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias.

Why Most Published Research Findings Are False

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias.

Essay

Much animal research into potential treatments for humans is wasted because it is poorly conducted and not evaluated through systematic reviews.

Is there evidence for a quality problem?

Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality.

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomisation</td>
<td>20%</td>
<td>53%</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>26%</td>
<td>54%</td>
</tr>
<tr>
<td>Blinded outcome assessment</td>
<td>28%</td>
<td>48%</td>
</tr>
<tr>
<td>Co-morbidity</td>
<td>18%</td>
<td>48%</td>
</tr>
</tbody>
</table>

Publication bias is highly prevalent (present in the literature describing the efficacy of at least 16 of 18 interventions) and accounts for around 30% of the reported efficacy of candidate neuroprotective interventions.

The 'Design' checklist (see course material)

- Internal validity?
- Proper statistics?
- Did ischemia-induced or pharmacologically induced systemic alterations confound the results?
- Proper controls for experiments with genetically modified animals?
- Proper observation interval after induction of focal cerebral ischemia?
- Did the experimental design allow the establishment of a causal relationship between experimental manipulation and phenotype?
- Proper use and interpretation of cerebral blood flow measurement?
- Edema correction?

Design: Internal validity and bias

- **selection bias** (creating groups with different confounders; solved by randomization)
- **performance bias and detection bias** (investigators respectively treating or assessing more positively those subjects on the treatment arm; controlled by blinding interventions and outcome assessments);
- **attrition bias** (dropouts of subjects with a negative outcome not included in the final result)

Design: External validity and bias

- External validity is a matter of judgment that depends on the characteristics of the subjects included, the setting, the treatment regimens, and the outcomes assessed.
- Perhaps the most consistently cited problems in experimental research concerns the validity of extrapolating data from young, healthy animals to elderly patients with frequent comorbid conditions in human clinical trials.

Quality issues

The 'Design' checklist (see course material)
**Design: Internal validity and bias**

- **selection bias** (creating groups with different confounders; solved by randomization)
- **performance bias** and **detection bias** (investigators respectively treating or assessing more positively those subjects on the treatment arm; controlled by blinding interventions and outcome assessments);
- **attrition bias** (dropouts of subjects with a negative outcome not included in the final result)

**Design: Internal validity**

- Hypothesis formulated, statistical analysis prespecified?
- Inclusion and exclusion criteria preset?
- Randomization
- Allocation concealment
- Blinded assessment of outcome

**Design: Mortality, selection bias**

Mortality after 60 min MCAO (mouse)

Days post MCAO

**Design: Cave systemic confounders I - Immune system**

- Untreated
- Antibiotic

**Design: Cave systemic confounders II - Wasting**

- Untreated
- Antibiotic

Stroke (2004) 35:2-6
Design: Control groups

- Historic controls, different batches
- ‘Sham’ controls
- Control groups in knockout experiments

Design: Control groups in transgenic animals

C57Black/6

SV129

Design: External validity and bias

- External validity is a matter of judgment that depends on the characteristics of the subjects included, the setting, the treatment regimens, and the outcomes assessed.
- Perhaps the most consistently cited problems in experimental research concerns the validity of extrapolating data from young, healthy animals to elderly patients with frequent comorbid conditions in human clinical trials.

External validity: Monitoring of physiological parameters

Hypothermia but not the N-methyl-D-aspartate antagonist, MK-801, attenuates neuronal damage in gerbils subjected to transient global ischemia.


External validity: Monitoring of physiological parameters


External validity: Timing of therapy

(Barber et al. Stroke. 2004;35:1720-1725.)
NMDA receptor blockade fails to alter axonal injury in focal cerebral ischemia.


External validity: 
Brain protection vs neuroprotection

3 month old rat 
20 month old rat

Volatile anesthetics induce tolerance against focal cerebral ischemia in the rodent


Hypoxia induces stroke tolerance in the mouse, but ..

Duration of MCAO

The results were significant at $p < 0.05$

1. The p value of a significant test is the probability that the research results are due to chance.
2. A hypothesis accepted as significant at the alpha level of significance has the probability of 1 – alpha of being found significant in future replications of the experiment.
3. A hypothesis accepted as significant at the alpha level of significance has a probability of 1 – alpha of being true.
4. The size of p of the significance level of a result is an index of the importance or size of a difference or relation.
5. The probability of rejecting $H_0$ is alpha.

P<0.05:
If we were to repeat the analysis many times, using new data each time, and if the null hypothesis were really true, then on only 5% of those occasions would we (falsely) reject it.

Scientists take samples. They make inferences about the population. But often scientists want an ‘objective method’ to decide whether an observation from a sample justifies the acceptance of a hypothesis about the population.

In other words, scientists want to test hypotheses, and make decisions!
So some guys early in the 20th century (Fisher, Pearson, Neyman, et al.) established formal ways of statistically analyzing data.

Key to this concept is the Null-Hypothesis.

We want to test whether two samples come from the same population.

$H_0$: There is no difference
$H_1$: There is a difference

If a researcher has a theory that a certain treatment has an effect, his theory is supported by rejecting another hypothesis (that there is no effect).

Since we sample, we make errors!
Two hypotheses, two types of error:

Type I: Rejected $H_0$, but $H_0$ was true (= false positive)

Type II: Accepted $H_0$, but $H_1$ was true (= false negative)

Type I error quantified by alpha (very often set to 0.05)
Type II error quantified by beta (very often ignored, sometimes set to 0.2). Power of study = 1 - beta.
Significance testing: The concept of power (Neyman, Pearson, Cohen)

Accept Reject

Distribution under $H_0$

Distribution under $H_1$

Power is the conditional probability of accepting the alternative hypothesis when it is true.

Power = $1 - \beta$ (Type II error level)

Statistical power is analogous to the concept of resolving power in evaluating optical instruments.

Is low power only a problem when $H_0$ is not rejected?

Assume that it is known that Compound X is effective in decreasing infarct sizes after stroke (= Null is false)

An experimental study finds 30% smaller infarcts after Compound X treatment. $p=0.02$

Power was 50%.

The study is repeated. What is the probability to find that compound X is effective?

Significance testing: alpha, beta, power

<table>
<thead>
<tr>
<th>Result of Exp.</th>
<th>Hypothesis wrong (= $H_0$ true)</th>
<th>Hypothesis correct (= $H_0$ false)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-significant (i.e. negative) result (= Accept $H_0$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significant (i.e. positive) result (= Reject $H_0$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Of the 95 studies that result in a significant result ($p<0.05$), 45 are true null hypotheses and so are false alarms!

Likelihood that a significant result is ‘true’: 52%

Thoughts can be transmitted from subject to subject (ESP) ($p<0.05$)

An experiment is conducted to see whether thoughts can be transmitted from one subject to another. Subject A is presented with a shuffled deck of cards and tries to communicate to Subject B whether each card is red or black by thought alone. In the experiment, Subject B correctly gives the color of 33 cards. The null hypothesis is that no thought transmission takes place and Subject B is randomly guessing. The observation of 33 correct is significant with a (one-sided) $p$-value of 0.035. Would you now believe that Subject A can transmit her thoughts to Subject B?

We recognize that much stronger evidence would be required to reject a highly plausible NULL. This makes clear that the $p$-value cannot mean the same thing in all situations.
Screening for a disease

A very accurate test: 0.1% false positives (very good sensitivity) 0.1% false negatives (very good specificity)

You take the test, test comes out positive!

What is the probability that you have the disease?

Incidence of the disease is 1:50000

Test is false positive at a rate of 1:1000

i.e.: Probability that you have the disease is 1:50!

The NULL (you do not have the disease) started with a very low probability!

Sensitivity ~ Type I error
Specificity ~ Type II error
Incidence of disease ~ Rate of true hypotheses

It is wrong to interpret a P-value as the probability of the null hypothesis, because this fails to take account of the prior probability of H0.

After O'Hagan and Luoo

Replication!

‘Good evidence’ that a hypothesized effect is real comes from replication across multiple studies and cannot be inferred from the result of a single statistical test!

Don't look for magical alternatives to NHST. There is no objective ritual to find out whether a hypothesis is true.

Significance testing: alpha, beta, power

Which factors determine alpha and beta?

Four factors combine:
1. size of the population effect,
2. sample size,
3. variance of the data,
4. alpha (or beta) level.

How?

1. Increasing alpha increases power (=decreases ß), but also increases the risk to reject H0 when it is true
2. Increasing sample size decreases the standard error, thereby increasing power
3. The larger the difference between the parameters tested, the greater the power to detect it

How does increasing sample size affect the SD?

Does the researcher have control over the factors that determine power?

1. Sample size?
2. Alpha level (and thus ß)
3. Variance?
4. Population effect?

GPOWER 3: Freeware power-analysis
(http://wwwpsycho.uni-dueesseldorf.de/aap/projects/gpower)
Highly recommended further reading or browsing:


LL Harlow, SA Mulaik, JH Steiger (Eds.) What if there were no significance tests? Lawrence Erlbaum Associates, London 1997


http://www.bayesian-initiative.com/

GPOWER 3: Freeware for power-analysis

http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Reporting

Special Communication

Improving the Quality of Reporting of Randomized Controlled Trials

The CONSORT Statement


Reporting of data in experimental stroke research

A

B

C

D

Reporting of data in experimental stroke research

A

B

Score

Control

CompX

Score

Control

CompX

0

1

2

3

4
'Consort'-like statement for experimental stroke research?

Standard operating procedures (SOPs)

'Write down what you do, and do what you write down'

The primary purpose of an SOP in brain research is to guide and standardize working procedures in order to ensure data reliability and integrity.

SOPs should be written by the user.

It is crucial that students, technicians, and researchers read and follow the SOPs. Otherwise they will not only fail to achieve their goal, but also engender a false sense of security.

Standard operating procedures (SOPs)

Good scientific practice

What is GSP?

Who's responsibility is it?

Documentation/Storage, honesty, openness

What are the legal aspects?

What are the fundamentals?

Good scientific practice

Data storage
Lab files
Authorship
Double publication, plagiarism
Reference to work of others
Dealing with misconduct
Whistleblowing
Conflicts of interest

GRP/GLP guidelines - every institution has one, no one reads it

Have you read them?
Do you know the name of the ombudsman/woman of your institution?
A GLP ‘consensus statement’

Wrap Up

- Quality matters
- Before starting the actual experiments, consider sources of bias (randomization, concealment of treatment allocation, inclusion/exclusion criteria)
- Start with a clear hypothesis, and prespecified statistical tests
- Perform a priori power analysis, and plan accordingly
- Consider to set up SOPs
- Report all the relevant information (ideally even raw data), and use maximally informative graphs. Report exclusions.
- Know the basics of GLP. Act accordingly.